Defense Threat Reduction Agency
8725 John J. Kingman Road, MS
6201 Fort Belvoir, VA 22060-6201

DTRA-TR-16-35

TECHNICAL REPORT

WMD Intent Identification and Interaction Analysis Using the Dark Web

Distribution Statement A. Approved for public release; distribution is unlimited.

April 2016

HDTRA1-09-1-0058

Hsinchun Chen

Prepared by:
University of Arizona
Tucson, AZ 85721

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 00-04-2016	**2. REPORT TYPE** Technical	**3. DATES COVERED** *(From - To)* N/A

4. TITLE AND SUBTITLE
WMD Intent Identification and Interaction Analysis Using the Dark Web

5a. CONTRACT NUMBER
HDTRA1-09-1-0058

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Hsinchun Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Arizona
Tucson, AZ 85721

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Threat Reduction Agency
8725 John J. Kingman Road, STOP 6201
Fort Belvoir, VA 22060-6201

10. SPONSOR/MONITOR'S ACRONYM(S)
DTRA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DTRA-TR-16-35

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Statement A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our primary goal is to develop computational models and techniques for understanding the intent and interaction patterns among adversarial parties. The basic research questions are grounded on computational linguistics and social media analytics. We are leveraging our highly successful Dark Web project as our research testbed (for identifying target adversarial groups and participants) and using other data sets and languages to verify performance. Approach: Develop basic research in Sentiment and Affect Analysis (SAA) and Interaction Coherence Analysis (ICA) to assist cyber analysts and policy makers in examining multilingual online WMD related Dark Web topic, sentiment, intent, and interaction dynamics.

15. SUBJECT TERMS
Dark Web; computational linguistics; WMD; sentiment analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Paul Tandy
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	19	**19b. TELEPHONE NUMBER** *(include area code)* 703-767-4663

UNIT CONVERSION TABLE

U.S. customary units to and from international units of measurement[*]

U.S. Customary Units	Multiply by → ← Divide by[†]	International Units
Length/Area/Volume		
inch (in)	2.54×10^{-2}	meter (m)
foot (ft)	3.048×10^{-1}	meter (m)
yard (yd)	9.144×10^{-1}	meter (m)
mile (mi, international)	$1.609\ 344 \times 10^{3}$	meter (m)
mile (nmi, nautical, U.S.)	1.852×10^{3}	meter (m)
barn (b)	1×10^{-28}	square meter (m^2)
gallon (gal, U.S. liquid)	$3.785\ 412 \times 10^{-3}$	cubic meter (m^3)
cubic foot (ft^3)	$2.831\ 685 \times 10^{-2}$	cubic meter (m^3)
Mass/Density		
pound (lb)	$4.535\ 924 \times 10^{-1}$	kilogram (kg)
unified atomic mass unit (amu)	$1.660\ 539 \times 10^{-27}$	kilogram (kg)
pound-mass per cubic foot (lb ft^{-3})	$1.601\ 846 \times 10^{1}$	kilogram per cubic meter (kg m^{-3})
pound-force (lbf avoirdupois)	4.448 222	newton (N)
Energy/Work/Power		
electron volt (eV)	$1.602\ 177 \times 10^{-19}$	joule (J)
erg	1×10^{-7}	joule (J)
kiloton (kt) (TNT equivalent)	4.184×10^{12}	joule (J)
British thermal unit (Btu) (thermochemical)	$1.054\ 350 \times 10^{3}$	joule (J)
foot-pound-force (ft lbf)	1.355 818	joule (J)
calorie (cal) (thermochemical)	4.184	joule (J)
Pressure		
atmosphere (atm)	$1.013\ 250 \times 10^{5}$	pascal (Pa)
pound force per square inch (psi)	$6.984\ 757 \times 10^{3}$	pascal (Pa)
Temperature		
degree Fahrenheit ($^{\mathrm{o}}$F)	[T($^{\mathrm{o}}$F) − 32]/1.8	degree Celsius ($^{\mathrm{o}}$C)
degree Fahrenheit ($^{\mathrm{o}}$F)	[T($^{\mathrm{o}}$F) + 459.67]/1.8	kelvin (K)
Radiation		
curie (Ci) [activity of radionuclides]	3.7×10^{10}	per second (s^{-1}) [becquerel (Bq)]
roentgen (R) [air exposure]	$2.579\ 760 \times 10^{-4}$	coulomb per kilogram (C kg^{-1})
rad [absorbed dose]	1×10^{-2}	joule per kilogram (J kg^{-1}) [gray (Gy)]
rem [equivalent and effective dose]	1×10^{-2}	joule per kilogram (J kg^{-1}) [sievert (Sv)]

[*] Specific details regarding the implementation of SI units may be viewed at http://www.bipm.org/en/si/.

[†] Multiply the U.S. customary unit by the factor to get the international unit. Divide the international unit by the factor to get the U.S. customary unit.

2015-11-16

Please answer all sections of the document. You are welcome to use figures and tables to complement or enhance the text. For annual reports, please only describe work for the period of performance. For final reports, please describe the comprehensive effort.

Grant/Award #:	**HDTRA1-09-1-0058**
PI Name:	**Hsinchun Chen**
Organization/Institution:	**University of Arizona**
Project Title:	**WMD Intent Identification and Interaction Analysis Using the Dark Web**
Report Period:	**Final Report, 2013-2014**

What are the major goals of the project?

List the major goals of the project as stated in the approved application or as approved by the agency. If the application lists milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion. Generally, the goals will not change from one reporting period to the next. However, if the awarding agency approved changes to the goals during the reporting period, list the revised goals and objectives. Also explain any significant changes in approach or methods from the agency approved application or plan.

Our primary goal is to develop computational models and techniques for understanding the intent and interaction patterns among adversarial parties. The basic research questions are grounded on computational linguistics and social media analytics. We are leveraging our highly successful Dark Web project as our research testbed (for identifying target adversarial groups and participants) and using other data sets and languages to verify performance.

- Approach: Develop basic research in Sentiment and Affect Analysis (SAA) and Interaction Coherence Analysis (ICA) to assist cyber analysts and policy makers in examining multilingual online WMD related Dark Web topic, sentiment, intent, and interaction dynamics.
- Objective: To investigate the WMD related capability, accessibility, and intent of potential extremist or terrorist groups using open source Internet content collected through our ongoing Dark Web project.
- Metrics: Evaluate and validate the proposed technologies based on the Dark Web corpora of extremist group web forums in English, Arabic and other languages, as well as on other forums and data sets.
- Status of effort: In this project period, we have continued to develop approaches to sentiment analysis and to investigate other methods that may lead to improvement in the identification and understanding of intent and capability.

What was accomplished under these goals?
For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results, including major findings, developments, or conclusions (both positive and negative); and 4) key outcomes or other achievements. Include a discussion of stated goals not met. As the project progresses, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

I. Major Activities

Early work in the project was focused on developing the initial research framework, shown in Figure 1, below.

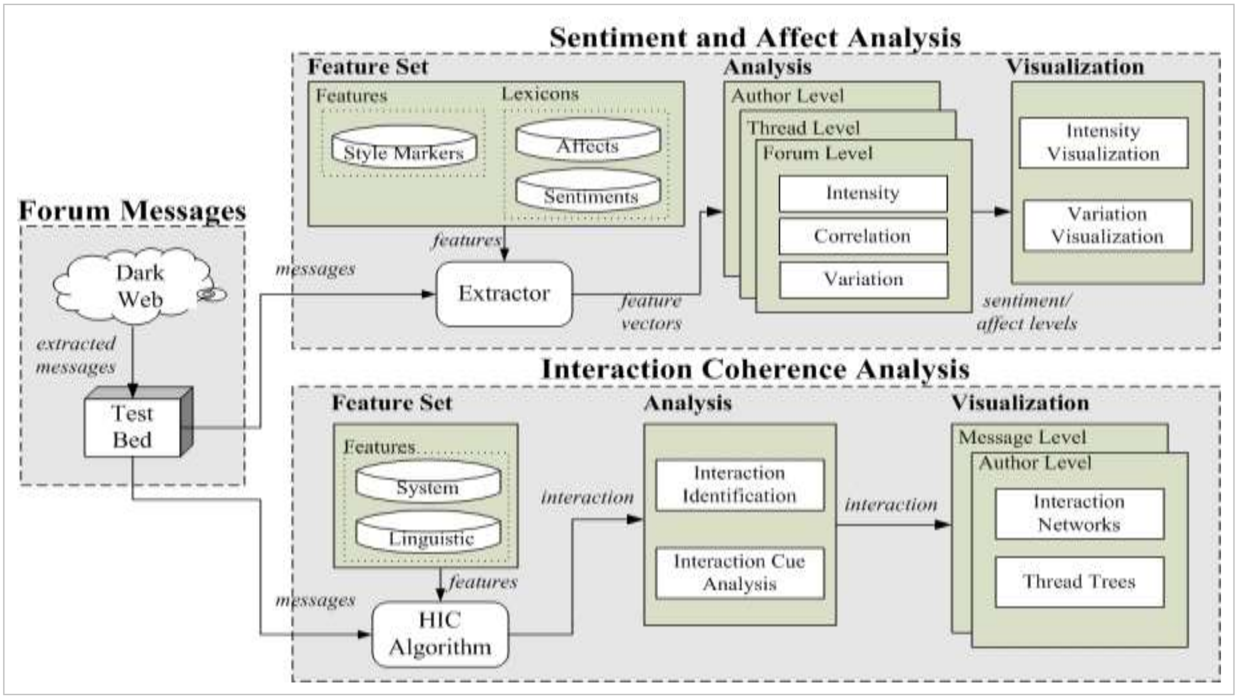

Figure 1. The original proposed research framework showing the development of the Dark Web test bed, feature-based sentiment and affect analysis of test bed messages, and feature-based interaction coherence analysis for network analysis.

In the ensuing periods, we

- Actively investigated methods for improving text sentiment analysis and developed a new feature selection technique, Feature Relation Network, which significantly improved over previous methods.
- Studied violent topic diffusion in web forums to identify the exposure process of extreme opinions through the novel adoption of the epidemic “SIR” (Susceptible-Infected-Recovered) model. We applied this model to a major Muslim forum that expounded extreme ideology.
- Expanded data collection methods by experimenting with developing a video portal. This required us to learn to develop spiders to query video-sharing websites using sets of keywords predefined by Dark Web domain experts, and to experiment with new video classification methods in order to correctly identify candidate videos.

- Devised and experimented with different sentiment classifiers across Dark Web forums to help address issues of scalability and minimizing training data requirements.
- Experimented with different methods for examining the relationships between extremist organizations.

During this last report period, we focused on completing and publishing results from our examination of the interaction dynamics and interconnectedness of virtual and real social movement organizations (SMOs). We also began to evaluate our techniques on a broader range of data sets; studies completed during this period involved testing our data collection and analysis methods in the domains of health information and stock prediction.

SMOs - Website Collection and Link Analysis

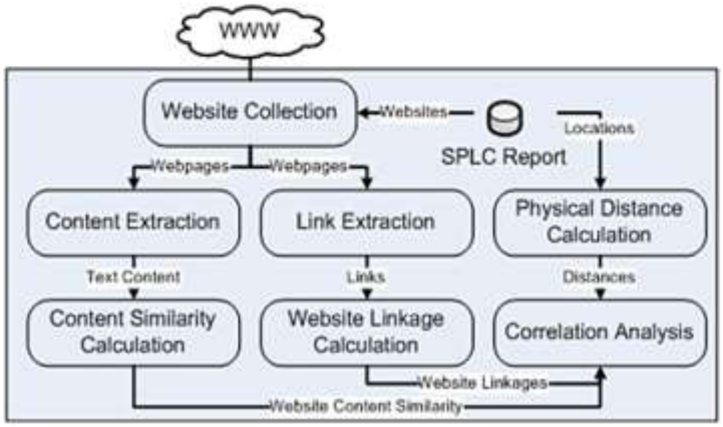

Figure 2. Our research design consisted of a multistage process. First, websites identified in the Southern Poverty Law Center (2009) report were collected using an automated spider. Next, the content and hyperlinks within each web document were extracted and processed for analysis. Finally, the two virtual attributes were correlated with real-world physical distance between groups.

This study was motivated by two questions:

(1) How do the virtual manifestations of social movement groups relate to their real-world presences?

(2) Does a social movement's ideological identity have an effect on the relationships between the virtual and real communities?

The data collection was focused on patriot groups and hate movements as identified by the Southern Poverty Law Center, an established legal advocacy group that tracks many American groups that advance socially deviant ideals. Founded in 1971 during the American civil rights movement, the SPLC combats hate, bigotry, and social injustices. It has routinely filed successful lawsuits against violent, socially extreme groups who seek to discriminate and exploit vulnerable members of society. Further, the SPLC is often recognized as a prominent force in continuing the fight against institutional racism. The SPLC releases a quarterly publication listing all domestic groups they track, along with additional commentary and information concerning studied organizations. Data on the social movements were acquired from the SPLC Spring 2009 Intelligence Report, which is comprised of a collection of 771 groups promoting extreme social perspectives through real-world activism in the United States and who also maintain strong virtual presences and participate in virtual activism.

We collected webpages from the virtual communities of social movement groups using a spider.

Table 1 summarizes the SPLC data set according to group ideological affiliations and categorizations. All groups belong to one of two subsets: hate groups or patriot groups.

Table 1. Southern Poverty Law Center Spring 2009 Intelligence Report data set. Ideological

Class	Subclasses	Description	Total no. of groups	Total no. of groups collected
Patriot	militia, media, ministries, publishing, support, vendor, political/citizen groups, and sovereign/common law/jural	opposed to the “new world order” or advocates extreme antigovernment doctrines	141	29
Hate	Ku Klux Klan, neo-Nazi,White nationalist, racist skinhead, Christian identity, neo-Confederate, Black separatist, and general hate	advocates extreme prejudice or violence on the basis of race or religion	630	74

The “Patriot” groups in general advocate extreme antigovernment doctrines. We collected pages from 29 of the 141 groups. The “Hate” groups advocate extreme prejudice and/or violence on the basis of race or religion; we collected pages from 74 of the 630 groups. For each organization, the top-three domain levels were collected as a representative sample of the community. All web pages were collected in a short time span to take a snapshot of the virtual communities. This collection of websites was then run through a series of analyses to better understand the relationships between the real world and virtual space.

We analyzed the collection in three dimensions. Hyperlinks were analyzed to measure the intensity of the virtual relationships established between any two groups in the data set. Content analysis provided a method for comparing the perspectives advanced between different groups. Virtual linkage intensities, content similarities, and geographical distances of the social movement groups were compared through several correlation analyses.

Website Collection

Web pages from the virtual communities of social movement groups were collected in an automated fashion using a spidering system. The top-three domain levels were collected as a representative sample of the community. All web pages were collected in a short time span to take a snapshot of the virtual communities. This collection of websites will be run through a series of analyses to better understand the relationships between the real world and virtual space.

Link Analysis

Hyperlinks were analyzed to measure the intensity of the virtual relationships established between any two groups in the data set. The amount of linkage (i.e., hyperlinks between groups between two groups) was defined as the total number of links from all pages within either website domain pointing to the other domain. Strong linkages between groups may be indicators of collaboration in the real world while weak linkages may be inferred as sparse or nonexistence of collaboration. We extracted hyperlinks from the collected web pages of each website to represent the linkages a virtual community maintains to other websites.

$$\textit{Website Linkage}_{(i,j)}=\textit{Links}_{(i\rightarrow j)}+\textit{Links}_{(j\rightarrow i)}.$$
$$\textit{Links}_{(i\rightarrow j)} = \textit{Number of hyperlinks from Website 1 pointing to Website 2.}$$
$$\textit{Links}_{(j\rightarrow i)} = \textit{Number of hyperlinks from Website 2 pointing toWebsite 1.}$$

Content Analysis

Content analysis provided a method for comparing the perspectives advanced between different groups. The textual content of websites was extracted after HTML code and function words were filtered. Websites with sparse content were excluded from analysis because they would unfairly skew the correlation analyses. To compute content similarity, word vectors were created from each group's website text to represent the content of their discussion and social position. The similarity of discussions and content publication between two groups within the data set was defined as the cosine similarity of their website word vectors. Stronger similarities may be an indicator of shared social perspectives and ideologies. Weaker similarities suggest a greater difference in ideals and focus.

Geographical Distance Calculation

Utilizing the real-world locations of the headquarters of each group, the geographical distance between all pairs of groups was calculated. Locations of group headquarters provided by the SPLC were transformed into latitude and longitude coordinates. Geographical distance was defined as the Euclidean distance between the two latitude and longitude coordinates. The amount of distance between two groups may have implications on relationships existing within the virtual space.

$$\textit{Geographical distance}_{(1,2)\,=}\sqrt{[(X_1 - X_2)^2 + (Y_1 - Y_2)^2]},$$

where X_1 is Group 1 latitude, Y_1 is Group 1 longitude, X_2 is Group 2 latitude, and Y_2 is Group 2 longitude.

Correlation Analysis

Virtual linkage intensities, content similarities, and geographical distances of the social movement groups were compared through several correlation analyses. Geographic distance is used as a real-world attribute to measure differences between the virtual aspects of groups. In essence, distance may have an influence upon the strength of relationships between the virtual representations of social movement groups. Correlations were performed at varying levels of ideological homogeneity as defined by SPLC categorizations. Analysis at multiple levels allows us to measure whether the ideologies of social movement groups affect aspects of their virtual communities. Three levels of ideological homogeneity were considered: the overall collection, he patriot- and hate-group ideological level, and the patriot and hate-subclass ideological level. Analysis at varying levels of ideological homogeneity is intended to reveal if the similarity of ideologies between groups affects certain aspects of their communities and relationships in the virtual space. In addition, differences may be revealed in how certain types of groups utilize virtual space. A log transform was applied to the linkage intensities prior to analysis and also to geographical distances when correlated with linkage. The log transform was implemented to reduce statistical skewness and to provide a more fair representation of the relationships between the virtual and real measures.

Research Hypotheses

Based on the literature, several expectations were established regarding the relationship between linkage intensity, content similarity, and geographical distance. We believe that the real-world aspect we observe, which is each group's geographic location, will have an influence on the virtual behaviors of groups. Previous literature has suggested that certain social movement groups are often founded within specific geographical locations, which would inherently place the group among others that share similar social perspectives (Dyke & Soule, 2002; McCann, 2009). Thus, groups geographically near each other may share social perspectives and subsequently may have stronger linkages and content similarities. In addition, as ideological homogeneity becomes more similar, stronger relationships are expected to form between the real and virtual aspects due to a higher similarity in social perspectives. The following research hypotheses were developed:

H1: A significant correlation will be observed between the physical distance between groups and their virtual interconnectedness.

H1a: A significant correlation will be observed between the physical distance between SMOs and their virtual linkage intensity.

H1b: A significant correlation will be observed between the physical distance between SMOs and their virtual content similarity.

H2: A significant correlation will be observed between content similarity and virtual linkage intensity among the social movement groups.

H3: Correlation analyses at varying degrees of ideological homogeneity among the groups will reveal distinctive relationships between the virtual aspects of the groups and their physical proximity.

H3a: The correlation observed will increase in significance as ideological homogeneity increases among the groups included in the analysis.

II. Specific Objectives

Our objectives for this reporting period were to test our methods against a broader range of data sets. Data were collected from social movement organizations as described above.

III. Significant Results

SMOs - Results & Discussion

In our work on SMOs, as hypothesized, different levels of analysis produced unique correlation analysis results. The results are summarized in Table 2.

Table 2. Correlation analysis results

Correlation analysis results		**Linkage intensity**	**Physical distance**
All groups	Linkage intensity	1	−.07983
	Content similarity	.00817	−.14658**
Patriot groups	Linkage intensity	1	.10477
	Content similarity	−.00238	.16441
Hate groups	Linkage intensity	1	−.11824
	Content similarity	.01295	−.17526**
Patriot–within subclass	Linkage intensity	1	.10816
	Content similarity	.06384	.43886*
Hate–within subclass	Linkage intensity	1	−.23593**
	Content similarity	.05622	−.18753**

Note. *p < .1. ***p* < .05, two-tailed test.

The third group of analyses focused on the relationships that groups held with other members of their own ideological subclass. The correlation between patriot-group geographical distance and content similarity is positive and significant; the results suggest that there are patriot groups dispersed across the United States with similar social perspectives and ideologies. This finding also supports H1b, H3, and H3A. Linkage intensity also was related to geographical distance in a positive direction, although at levels that were not statistically significant. Patriot groups link to other groups that are geographically far apart, perhaps to span distances. In addition, these groups may wish to connect with others that share similar social perspectives, but they may avoid forming ties with physically near groups. This behavior may indicate competition between groups to recruit future members, as observed in Zhou et al. (2005). Without linking to others who are physically near, a group has the potential to recruit more individuals from the local population.

Analysis of hate groups within subclasses also revealed interesting relationships. Content similarity and geographical distance remain strongly correlated in a negative direction for hate groups. Groups closer to one another in the physical world were found to have more similar content. The relationship between linkage and geographical distance is significant when the analysis is restricted to ideological subclasses. Previous literature has found that hate groups coordinate real-world events to unite members and promote similar messages. They do so to recruit new members, promote propaganda, and attract publicity (Brower, 2009). This may be interpreted as evidence that hate groups use the Internet to coordinate real-world activism, and their ideologies are more geographically limited in context than patriot groups. This supports H1a, H1b, H3, and H3A.

Counterintuitively, virtual linkage intensity and content similarity were not found to be significantly correlated at any level of ideological homogeneity. Intuition would suggest that significant correlations would be discovered between the two virtual aspects of the social movement groups. Upon further investigation of some of the relationships existing within our data set, we found that interconnected groups may link to each other because they share similar social perspectives, but that their actual content discussed different, yet related, concepts. For example, two heavily interconnected groups, the “Church of the sons of YHVH” and “Aryan Nations,” both share White nationalist ideology. However, one group concentrates on promoting anti-Jewish sentiment while the other group’s content tends to cover more political concepts and events.

To better understand the social movement groups, we further investigated our analysis results through illustrative examples. In our examples, the markers on the map represent different groups. Groups of interest are denoted with green markers and are numbered. Pictures of their websites also are included. Red markers refer to hate groups, and blue markers indicate patriot groups. Lines between groups are illustrations of discussed relationships, with line thickness denoting the relationship’s strength. Patriot groups have significant correlations concerning content similarity and geographical distance while hate groups have significant relationships concerning both virtual aspects when correlated with physical proximity.

Figure 3 demonstrates the content similarity of the American Patriot Friends Network (AFPN) within the patriot-group ideological level. The APFN characterizes the “American patriot” as an individual who upholds the U.S. Constitution and is skeptical of the American government operating within the legal bounds of the Constitution. It has strong content similarities with other groups who share constitutionalist perspectives, and weak linkages with groups who may focus their activism on other causes. One strong instance of content similarity exists between the APFN (A) and the Conservative USA group (B). Both groups share similar perspectives in referencing the Constitution as the highest authority in the United States. This behavior is again observed between the APFN and the Lawful Path group (C). The Lawful Path group is dedicated to ensuring that the U.S. government abides by the Constitution, an interest shared with the AFPN. However, the APFN has much weaker content similarity with groups who refer to constitutional authority, but actually focus on different topics. For example, the Liberty Gun Rights (D) group focuses its efforts on advancing gun rights, and uses the Constitution only as a minor justification for gun ownership should be unregulated; the APFN has weak content similarity with this group.

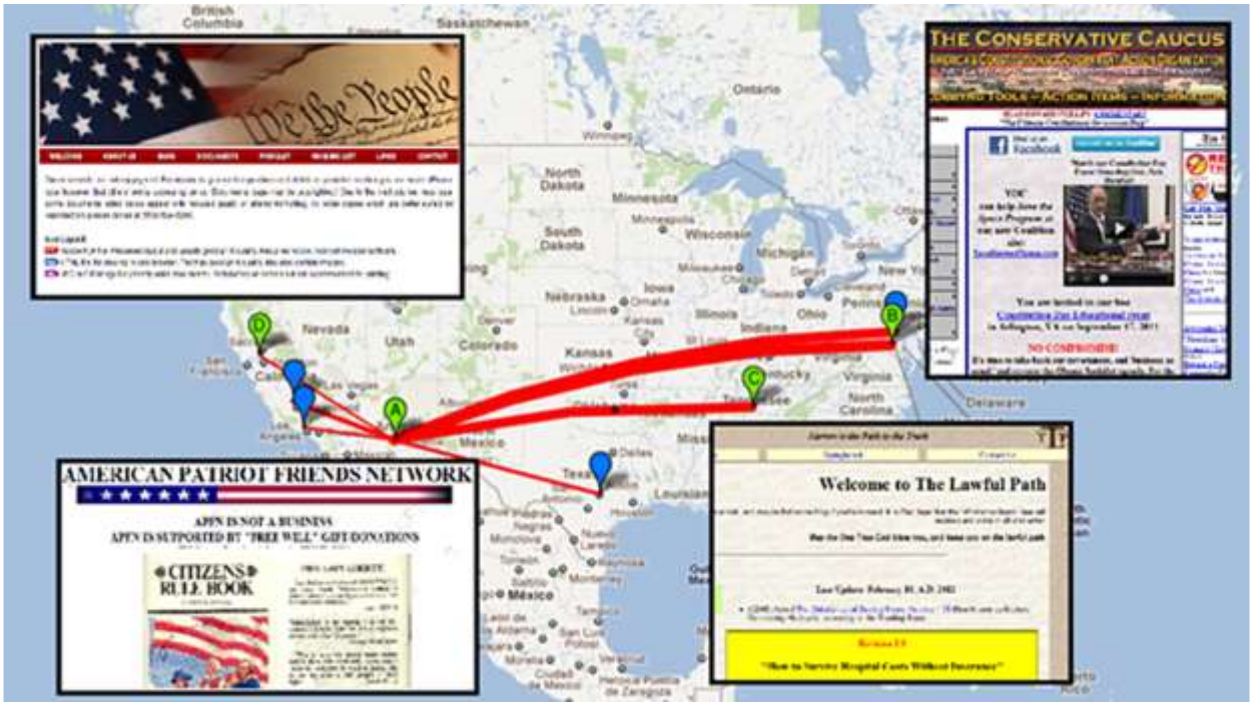
THE CONSERVATIVE CAUCUS
AMERICAN PATRIOT FRIENDS NETWORK
APFN IS NOT A BUSINESS
APFN IS SUPPORTED BY "FREE WILL" GIFT DONATIONS
CITIZENS RULE BOOK
Welcome to The Lawful Path
A
B
C
D

Figure 3. An example of virtual content similarity behavior between various patriot groups. The American Friends Patriot Network (A) has strong similarity with other constitutionalist groups such as Conservative USA (B). Both groups discuss political news and events from the perspective of constitutionalism. However, they have weaker relationships with groups who may rely on constitutional ideals but have different goals, such as the pro-gun Liberty Gun Rights (D).

An example of content similarity within the hate-group ideological level can be seen in Figure 4. The League of the South (A), in Killen, Alabama, has stronger content similarity with groups who are nearby as opposed to groups who are physically distant. Furthermore, the League of the South has relationships with nearby groups who take pride in American CivilWar Confederate perspectives; the strongest content similarities are with the Virginia League of the South (B), the Louisiana League of the South (C), and the Florida League of the South (D). All groups are local chapters of an umbrella organization. This type of relationship between groups is expected, as previous literature has stated that hate groups are generally geographically constrained to regions holding conservative social perspectives (Fording & Cotter, 2007). The League of the South is limited by this geographical constraint, and thus chapters of the organization are necessarily physically near to each other. Conversely, the Alabama League of the South has weak relationships with other American nationalist organizations that do not hold a Confederate identity.

Figure 4 demonstrates some of the observed virtual linkage behavior within the hate-group ideological level. The Covenant People's Ministry (CPM), in Brooks, Georgia, has stronger virtual linkage with groups that are physically nearby while possessing weak virtual linkage with groups who are further away. Within this relationship between linkage and distance, CPM (A) has strong linkage intensities with other groups that claim to be upholding Christian values while propagating White supremacist agendas, such as the Church of the Sons of YHVH (B) and StormFront (C). Christian groups that lacked a White supremacist ideology, such as the America's Promise group (D), had much weaker linkage with CPM.

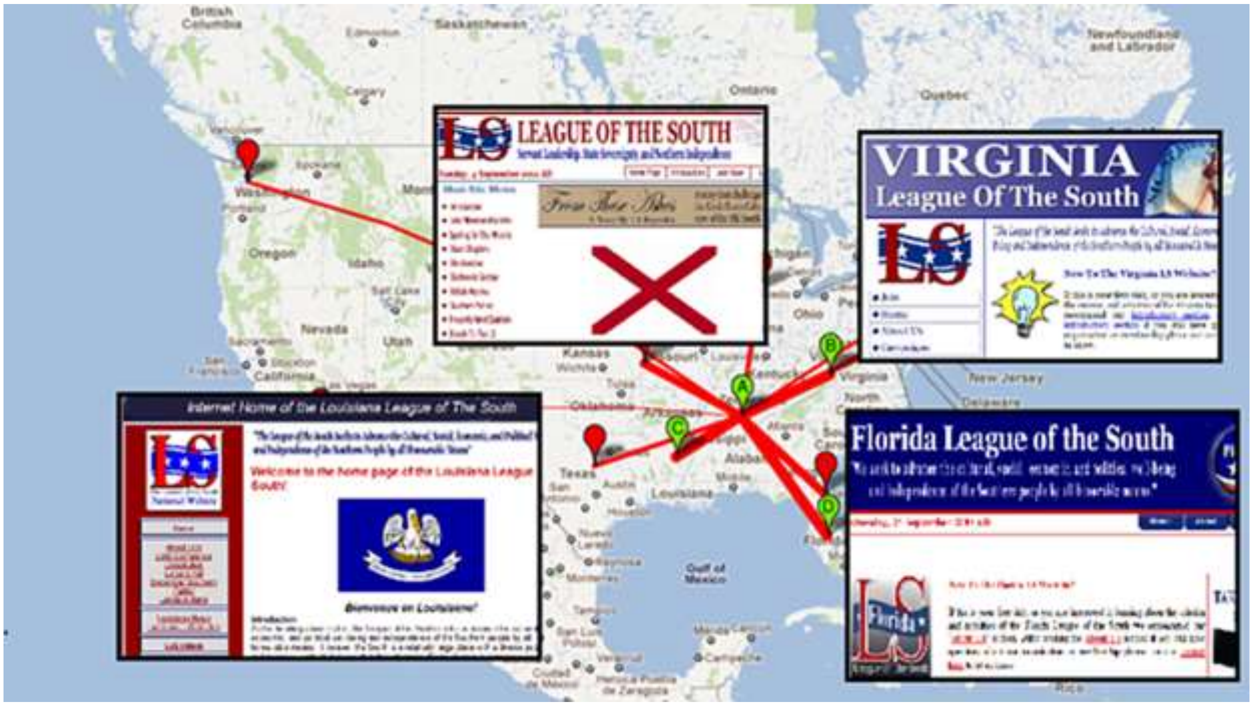

Figure 4. Content similarity behavior between hate groups. Social movement groups belonging to the umbrella organization "League of the South" all share similar web content.A"League of the South" banner can be found on the web pages of all affiliated groups, further demonstrating close collaboration among involved groups.

Patriot and hate groups exhibited distinctive characteristics in their virtual behaviors; patriot groups are more widely dispersed across the United States while hate groups tend to be clustered in specific geographic regions. Our results indicated that patriot groups dispersed across the United States share similar social positions, but that they may not necessarily be interconnected. This may be due to findings outlined in previous studies that patriot groups discuss topics concerning social instability, but that their focus is on a geographically local level or a specific aspect of societal structure (Dyke & Soule, 2002). In addition, the results showed that patriot groups are more likely to have weaker content similarities with other groups who are nearby a opposed to those who are geographically far. Physically near groups who share a local population from which to recruit new members may purposely focus their content on differing social issues to avoid competing over future members (Zhou et al., 2005). Patriot groups may have stronger content similarities with geographically remote groups due to this lack of competition; however, given their local focus, there may not be an interest in connecting to far-away groups despite shared ideologies.

Hate groups located near one another geographically were tightly interconnected virtually. Relationships are particularly significant when analysis is restricted to hate groups of the same subclass. This may indicate coordination between physically near groups in conducting real-world activism driven by mutually held ideologies, supporting the conclusions of previous studies. Virtual linkage and content similarity were not found to be significantly correlated at any level of ideological homogeneity. Manual scrutiny of content revealed that many groups discuss similar topics, but in different contexts, perhaps explaining the lack of correlation.

This research provided a framework for future analyses of the relationship between virtual and real-world aspects. The correlation of linkage intensity and content similarity with geographical distance is unique to this study. In addition, this research provided insights into social movements; in particular, our results revealed how social movements' online behaviors are influenced by real-world social and geographical positions.

IV. Training and Dissemination

What opportunities for training and professional development has the project provided?
If the research is not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report." Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

We are grateful for the encouragement we received from DTRA to involve students in all aspects of the research as well as support their professional development. The Ph.D. students designed experiments, devised data collection plans, and trained new and junior students. The undergraduate students typically helped with system development and administration, and worked their way up to helping find relevant papers for literature reviews.

Beginning graduate students took on tasks related to database administration, or spidering (data collection), for example, and usually worked closely with a more senior student.

How have the results been disseminated to communities of interest?
If there is nothing significant to report during this reporting period, state “Nothing to Report.” Describe how the results have been disseminated to communities of interest. Include any outreach activities that have been undertaken to reach members of communities who are not usually aware of these research activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

We again attended the IEEE Intelligence and Security Informatics conference, which included a mix of faculty and practitioners from a variety of institutions, including universities, national labs, and corporations. We also continue to publish in top-tier journals so as to reach the largest number of relevant audience members.

We find our work to be of interest to computer and information scientists, network engineers, social scientists, and government practitioners.

What do you plan to do during the next reporting period to accomplish the goals?
If there are no changes to the agency-approved application or plan for this effort, state “No Change.” Describe briefly what you plan to do during the next reporting period to accomplish the goals and objectives.

N/A - this is the last reporting period.

V. Personnel Associated with the Research Effort (all periods):

- Research Staff:
  - Hsinchun Chen, PI
  - Catherine A. Larson, Co-PI
  - Andrew Pressman, Research & systems technician
- Ph.D. Students:
  - Victor Benjamin, Ph.D. student
  - Yukai Lin, Ph.D. student
  - Xiao Liu, Ph.D. student
  - Shan Jiang, Ph.D. Student
  - Theodore Elhourani, Ph.D. student (graduated)
  - David Zimbra, Ph.D. student (graduated)
  - Qiang Gao, Ph.D. student (graduated)
  - Shuo Zeng, Ph.D. student (changed programs)
  - Huanchen Zhang, Ph.D. student (left program)
- Masters and Undergraduate Students:
  - Chung-Huan Hsieh, Masters student (graduated)
  - Chung-Ting Shing, Masters student (graduated)
  - Joe Parsons, Masters student (graduated)
  - Edward Huang, Masters student (graduated)
  - David Ware, Undergraduate student (graduated)
  - Joshua Clusin, Undergraduate student (graduated)
  - Tim Wang, Undergraduate student (graduated)

- Collaborators:
  - Dr. Hsin-Min Lu, NTU faculty member
  - Dr. Shu-Shing Li, NTU faculty member

VI. Publications

This reporting period:

(1) D. Zimbra, H. Chen, and R. Lusch. Stakeholder Analyses of Firm-Related Web Forums: Applications in Stock Return Prediction (2014). *ACM Transactions on Management Information Systems.* Forthcoming; accepted September 2014. Acknowledgement of federal support: Yes

(2) V. Benjamin, H. Chen, and D. Zimbra (2014). "Bridging the Virtual and Real: The Relationship Between Web Content, Linkage, and Geographical Proximity of Social Movements." *Journal of the Association for Information Science and Technology*. Forthcoming; first published online April 28, 2014. Acknowledgement of federal support: Yes

(3) J. Chuang, O. Hsiao, P.-L. Wu, J. Chen, X. Liu, H. De La Cruz, S.-H. Li, and H. Chen (2014). "DiabeticLink: An Integrated and Intelligent Cyber-Enabled Health Social Platform," in X. Zheng et al. (Eds.): International Conference on Smart Health (ICSH), Lecture Notes in Computer Science (LNCS) v. 8549, pgs. 63–74. Acknowledgement of federal support: Yes

(4) Y.K. Lin, H. Chen, R. Brown, S.-H. Li, H.-J. Yang (2014). "Time-to-Event Predictive Modeling for Chronic Conditions Using Electronic Health Records." *IEEE Intelligent Systems* 29(3): May-June, pgs. 14-20. Acknowledgement of federal support: Yes

(5) Y.K. Lin, H. Chen, R. Brown (2013). "MedTime: A temporal information extraction system for clinical narratives." *Journal of Biomedical Informatics* 46, s20-s28. Acknowledgement of federal support: Yes

Previous reporting periods:

(6) Benjamin, Victor; Chung, Wingyan; Abbasi, Ahmed; Chuang, Joshua; Larson, Catherine A.; Chen, Hsinchun. (2013) "Evaluating Text Visualization: An Experiment in Authorship Analysis." In Proceedings of the IEEE Intelligence and Security Informatics Conference (June 4-7, Seattle, Washington,). Published. Acknowledgement of federal support: Yes

(7) Benjamin, Victor; Chen, Hsinchun. (2013). "Machine Learning for Attack Vector Identification in Malicious Source Code." In Proceedings of the IEEE Intelligence and Security Informatics Conference (June 4-7, Seattle, Washington,). Published. Acknowledgement of federal support: Yes

(8) Chen, Hsinchun. (2012). *Dark Web: Exploring and Data Mining the Dark Side of the Web.* Springer. Acknowledgement of federal support: Yes

(9) Hsinchun Chen, Roger H.L. Chiang, Veda C. Storey (2012). Business intelligence and Analytics: From Big Data to Big Impact. 36. (4). *MIS Quarterly* (Special Issue: Business Intelligence Research), 36. 1165. Published. Acknowledgement of federal support: Yes.

(10) Fu, Tianjun; Abbasi, Ahmed; Zeng, Daniel; Chen, Hsinchun (2012). "Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers." *ACM Transactions on*

Information Systems 30(4), November, Article 24.

(11) Zimbra, David; and Chen, Hsinchun. (2012). Scalable Sentiment Classification Across Multiple Dark Web Forums. In Proceedings of the IEEE Intelligence and Security Informatics Conference (June 11-14, Washington, D.C.). Published. Acknowledgement of federal support: Yes

(12) Benjamin, Victor; and Chen, Hsinchun (2012). Securing Cyberspace: Identifying Key Actors in Hacker Communities. *In* Proceedings of the IEEE Intelligence and Security Informatics Conference (June 11-14, Washington, D.C.). Published. Acknowledgement of federal support: Yes.

(13) Woo, Jihung; Chen, Hsinchun (2102). "An event-driven SIR model for topic diffusion in web forums. *In* Proceedings of the IEEE Intelligence and Security Informatics Conference (June 11-14, Washington, D.C.). Published. Acknowledgement of federal support: Yes.

(14) Yang, Ming; Chen, Hsinchun (2012). "Partially supervised learning for radical opinion identification in hate group web forums." *In* Proceedings of the IEEE Intelligence and Security Informatics Conference (June 11-14, Washington, D.C.). Published. Acknowledgement of federal support: Yes.

(15) Zimbra, David. Stakeholder and Sentiment Analysis in Web Forums (2012). Dissertation Submitted to the Faculty of the Department of Management Information Systems In Partial Fulfillment of the Requirements For the Degree of Doctor of Philosophy, Graduate College, The University of Arizona. Deposited. Acknowledgement of federal support: Yes.

(16) Chen, Hsinchun. Smart Market and Money, *IEEE Intelligent Systems* 26(6), Nov-Dec 2011, pp. 82-84. Published. Acknowledgement of federal support: Yes.

(17) Chen, Hsinchun; Huang, Edward Chun-Neng; Lu, Hsin-Min; and Li, Shu-Shing (2011). AZ SmartStock: Stock Prediction with Targeted Sentiment and Life Support. IEEE *Intelligent Systems* 26(6), Nov-Dec, pp. 84-88. Published. Acknowledgement of federal support: Yes.

(18) Zimbra, David; and Chen, Hsinchun (2011). A Stakeholder Approach to Stock Prediction Using Finance Social Media. IEEE *Intelligent Systems* 26(6), Nov-Dec, pp. 88-92. Published. Acknowledgement of federal support: Yes.

(19) Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen (2011).. "Selecting Attributes for Sentiment Classification Using Feature Relation Networks." IEEE *Transactions on Knowledge and Data Engineering*. 23:3. 2011. Published.

(20) Hsinchun Chen, Dorothy Denning, Nancy Roberts, Catherine A. Larson, Ximing Yu and Chun-Neng Huang (2011). "The Dark Web Forum Portal: From Multi-lingual to Video." IEEE Intelligence and Security Informatics 2011 Conference. Published. Acknowledgement of Federal Support: yes

(21) Woo, Jiyung; Son, Jaebong, Chen, Hsinchun (2011). "An SIR model for violent topic diffusion in social media." IEEE Intelligence and Security Informatics 2011 Conference. Published. Acknowledgement of Federal Support: yes

(22) Hsinchun Chen. "From Terrorism Informatics to Dark Web Research." In Uffe Kock Wiil, ed. Counterterrorism and Open Source Intelligence." Lecture Notes in Social Networks, 2011, Volume 2, Part 3, Page 317--. Springer 2011. Published. Acknowledgement of Federal Support: yes

(23) Yulei Zhang; Ximing Yu; Yan Dang; Hsinchun Chen. “An Integrated Framework for Avatar Data Collection from the Virtual World.” IEEE *Intelligent Systems*, 25:6, 2010. Page 17-- . Published. Acknowledgement of Federal Support: Yes.

(24) Yulei Zhang, Shuo Zeng, Chun-Neng Huang, Li Fan, Ximing Yu, Yan Dang, Catherine A. Larson, Dorothy Denning, Nancy Roberts, and Hsinchun Chen. “Developing a Dark Web Collection and Infrastructure for Computational and Social Sciences.” IEEE Intelligence and Security Informatics 2011 Conference. IEEE, 2010. Published. Acknowledgement of Federal Support: yes.

(25) Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, and Jay F. Nunamaker, Jr. “Detecting Fake Websites: The Contribution of Statistical Learning Theory.” *MIS Quarterly*, 34:3, 2010. Page 435--. Published.

VII. Interactions/Transitions:

Dr. Hsinchun Chen has been appointed the Program Director for the National Science Foundation’s Smart and Connected Health Program beginning 2014.

VIII. New discoveries, inventions, or patent disclosures

None at this time.

IX. New Honors and Awards

None at this time.

DEPARTMENT OF DEFENSE

DEFENSE THREAT REDUCTION
AGENCY
8725 JOHN J. KINGMAN ROAD
STOP 6201
FORT BELVOIR, VA 22060
ATTN: P. TANDY

DEFENSE TECHNICAL
INFORMATION CENTER
8725 JOHN J. KINGMAN ROAD,
SUITE 0944
FT. BELVOIR, VA 22060-6201
ATTN: DTIC/OCA

DEPARTMENT OF DEFENSE CONTRACTORS

QUANTERION SOLUTIONS, INC.
1680 TEXAS STREET, SE
KIRTLAND AFB, NM 87117-5669
ATTN: DTRIAC